CONSERVING ENERGY IN A DATA PROCESSING NETWORK

BACKGROUND

1. Field of the Present Invention

The present invention generally relates to the field of network computing and more particularly to a method and system for reducing energy consumption in a server cluster by dynamically adjusting the operating frequency of selected server-network links.

2. History of Related Art

In the field of networked computing and data processing, server clusters are commonly used as a means of providing network services. A server cluster typically includes a set of server devices, each of which is capable of processing server requests. The cluster may include a request distributor that is configured to route incoming requests to an appropriate server in the server cluster for processing. Requests may be distributed to the individual servers based upon the current loading of the individual servers, the origin of the request, the location of the requested file or data, or other appropriate factors.

Server clusters are frequently arranged according to a switched configuration in which each server communicates with a central switch via a transmission medium such as twisted copper, fiber optic cable, or wirelessly transmitted electromagnetic waves. When the network parameters are configured, a transmission rate is established for each server-switch link based upon the bandwidth capabilities of the respective network interface cards and the transmission medium itself. Typically, the transmission rate for a given link is determined when the link is established and remains set during the link lifetime. Moreover, the transmission rate that is established is typically the highest possible transmission rate that both ends of the link can accommodate.

Maintaining the transmission rate of each network link at the highest possible value maximizes performance but only at the cost of increased power consumption. It is common knowledge that operating a network link at high frequency costs more than operating the same link at low frequency. Moreover, the additional cost incurred to operate the network links at high frequency often does not translate into correspondingly improved performance because the

data transmission rate may be limited by factors other than the physical bandwidth of the link between the server and switch.

The sum of the bandwidth of the individual server-switch links cannot exceed the bandwidth allocated to the server cluster as a whole. Thus, if a server cluster having an allocated bandwidth of 200 Megabits/second (Mbps) is supporting a total of 20 servers, each connected to a central switch with a 100 Mbps link, it is physically impossible for all of the links to operate at their maximum bandwidth simultaneously. Moreover, the connection between a remote client and the server cluster may represent a limit on the usable bandwidth of the server-switch link. If a client connects to the server cluster (and an individual server) over a 56 Kbps modem connection during a period when there is no other network traffic, the maximum bandwidth of the server-switch link that can be utilized to service the client request is 56 K. If the server-switch link is operating at 100 Mbps as an example, the bandwidth will be severely underutilized. It would therefore, be desirable to implement a method and system for dynamically conserve energy consumption in a data processing network by dynamically optimizing the operating frequencies of the server links in response to changing network conditions.

SUMMARY OF THE INVENTION

The problems identified above are in large part addressed by a data processing network and method in which the operating frequency of network links is adjusted dynamically to conserve energy consumption with a minimum of performance loss. When the maximum usable bandwidth of a server's network link is less than the current operating frequency of the link, the operating frequency of the server link is reduced. Similarly, if the maximum usable bandwidth of the link exceeds the current operating frequency, the operating frequency may be increased. In one embodiment, the data processing network includes a server cluster in which a set of server devices are connected to a central switch. The individual server-switch links may comply with an industry standard network configuration protocol such as Ethernet. Initially, the server-switch links may be established at the link's maximum operating frequency according to a negotiation process specified in a protocol such as IEEE 802.3. Periodically, thereafter, the server may determine that the current operating frequency of its link exceeds the capacity required to service

client requests while maintaining a desired level of performance. The server (or the switch) may then adjust the bandwidth of its link to operate at the lowest possible operating frequency required to accommodate the current loading. In this manner, the data processing network reduces power consumption by minimizing the operating frequency of its individual server links.

BRIEF DESCRIPTION OF THE DRAWINGS

Other objects and advantages of the invention will become apparent upon reading the following detailed description and upon reference to the accompanying drawings in which:

FIG 1 is a block diagram illustrating selected features of a data processing network;

FIG 2 is a block diagram illustrating additional detail of the data processing network of FIG 1;

FIG 3 is a block diagram illustrating additional detail of the network interface card of FIG 2; and

FIG 4 is a flow diagram illustrating operation of a server in the data processing network of FIG 1.

While the invention is susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that the drawings and detailed description presented herein are not intended to limit the invention to the particular embodiment disclosed, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the present invention as defined by the appended claims.

DETAILED DESCRIPTION OF THE INVENTION

Turning now to the drawings, FIG 1 is a block diagram of selected features of a data processing network 100 according to one embodiment of the present invention. In the depicted embodiment, data processing network 100 includes a server cluster 101 that is connected to a wide area network (WAN) 105 through an intermediate gateway 106. WAN 105 may include a multitude of various network devices including gateways, routers, hubs, and so forth as well as

one or more local area networks (LANs) all interconnected over a potentially wide-spread geographic area. WAN 105 may represent the Internet in one embodiment.

Server cluster 101 as depicted includes a central switch 110 that is connected to the gateway 106 via a network link 200. Cluster 101 further includes a plurality of servers, four of which are depicted in FIG 1 and indicated by reference numerals 111-1, 111-2, 111-3, and 111-4. Each server 111 is connected to switch 110 via a dedicated network link (reference numerals 211, 212, 213, and 214).

Server cluster 101 may service all requests to a single universal resource indicator (URI) on network 100. In this embodiment, client requests to the URI originating from anywhere within WAN 105 are routed to server cluster 101. Switch 110 typically includes a request distributor software module that is responsible for routing client requests to one of the servers 111 in cluster 101. The request distributor may incorporate any of a variety of distribution algorithms or processes to optimize the server cluster performance, minimize energy consumption, or achieve some other goal. Switch 110 may, for example, route requests to a server 111 based on factors such as the current loading of each server 111, the source of the client request, the requested content, or a combination thereof.

In one embodiment, network links 211, 212, 213, and 214 utilize the Ethernet protocol. In this embodiment, each server 111 includes an Ethernet compliant network interface card and switch 110 includes an Ethernet compliant port for each server 111. Referring to FIG 2, a block diagram illustrating additional detail of a switch 110 and one of the servers 111 is presented. Each server 111 includes a NIC 121 that connects to a corresponding port 131 in switch 110. In one embodiment suitable for use in the present invention, NIC 121 and each port 131 of switch 110 are capable of operating at various operating frequencies. In one embodiment, for example, NIC 121 and its corresponding port 131 are capable of supporting Ethernet links operating at 10 Mbps, 100 Mbps, and 1000 Mbps. Commercially available switches with such capability are represented by, for example, the 180 series of content-intelligent web switches from Alteon Web Systems (www.alteonwebsystems.com). Similarly, network interface cards such as the 10/100/1000 PCI-X Server NIC from 3Com may provide the ability to operate at different operating frequencies.

Servers 111 and switch 110 are configured to engage in a negotiation process to arrive at and agree upon an operating frequency for the corresponding link between them. In a

conventional server cluster configuration, this negotiation is performed only during link initialization and the negotiation outcome, including the link's operating frequency, remains constant as long as the link is present. Moreover, the operating frequency that the negotiation produces is typically the maximum operating frequency that the switch, server, and interconnecting medium can accommodate. The present invention contemplates a system and method for periodically modifying the operating frequencies of the various server-switch links in response to changing server cluster conditions to achieve a desirable level of cluster response performance while reducing the operating cost of the server cluster.

Turning now to FIG 3, a block diagram illustrating additional detail of a NIC 121 suitable for use in the present invention is depicted. The depicted embodiment of NIC 121 includes an embedded processor 140 that interfaces to a peripheral bus or local bus 144 of server 111. Bus 144 is typically implemented according to an industry standard bus protocol such as the Peripheral Components Interface (PCI) local bus as specified in *PCI Local Bus Specification 2.2* from the PCI Special Interest Group (www.pcisig.com). NIC 121 further includes buffer logic 141 connected to processor 140 that provides temporary storage for information received from and transmitted to network link 211.

A clock generator 142 provides the basic clocking signal 148 that drives buffer logic 211 and thereby establishes the operating frequency of network link 211. In the depicted embodiment, clock generator 142 is capable of providing clocking signal 148 at various frequencies controlled by the settings in a clock register 146. Clock register 146 is under the programmable control of processor 140. A memory 143 is accessible to processor 140 and buffer logic 141. Memory 140 may include volatile storage such as a conventional dynamic or static random access memory (DRAM or SRAM) array as well as persistent or non-volatile storage such as a flash memory card or other form of electrically erasable programmable read only memory (EEPROM).

Portions of the present invention may be implemented as a computer program product comprising a set of computer executable instructions stored on a computer readable medium. The computer readable medium in which the instructions are stored may include the volatile or non-volatile elements of memory 143. Alternatively, the instructions may be stored on a floppy diskette, hard disk, CD ROM, DVD, magnetic tape, or other suitable persistent storage facility.

NIC 121 includes software configured to perform a negotiation with switch 110 via the corresponding network link to establish the link's operating frequency. In an Ethernet embodiment of server cluster 101, the negotiation process software is typically compliant with the IEEE 802.3 standard, which is incorporated by reference herein. Ethernet compliant NIC's and switches typically include code that establishes the operating frequency of the network link. As indicated previously, this code is executed only when the link is established in a conventional server. NIC 121 and its corresponding port 131 according to the present invention, however, are both configured to invoke this negotiation process code periodically to modify the link operating frequency in response to changing conditions in the bandwidth utilization of the link.

Referring to FIG 4, a flow diagram illustrating a method of controlling the operating frequency of various links in a data processing network such as server cluster 101 is presented. When a server-switch link is first established, NIC 121 will initiate (block 402) a negotiation referred to herein as the original negotiation. Typically, the original negotiation establishes the maximum link operating frequency that the components can accommodate. Thereafter, NIC 121 monitors (block 404) the utilization of the network link between itself and its corresponding port. NIC 121 is configured to recognize periods of significant under-utilization of the server-switch bandwidth and to adjust the operating frequency of the link accordingly.

The link utilization monitored by NIC 121 represents the rate at which data is transmitted and/or received over the link. This utilization may be determined using a relatively simple link level routine in which a link utilization factor (also referred to herein as an effective data rate) is determined periodically. The routine would typically determine the volume of traffic transmitted and/or received over the link during a specified time period using an accumulator or other suitable mechanism. The specified time period may coincide with the periodic rate at which the link operating frequency is updated. If, for example, the link operating frequencies are to be modified, if needed, every ten minutes, the utilization factor may be determined by accumulating the number of bytes of link traffic over a ten minute period and dividing by 600 seconds to obtain a utilization rate in terms of bytes per second. The periodic intervals at which the link operating frequency is modified is preferably under the programmable control of switch 110 or server 111 such that the specified time period may be altered.

The effective transmission rate may be substantially less than the operating frequency of the network link. The bandwidth of the switch-gateway link 200 provides an upper limit on the

sum of the bandwidths of the individual server-switch links 211, 212, etc. The bandwidth needed for any individual server-switch link cannot exceed the bandwidth allocated to switch-gateway link 200. Moreover, the effective data transmission rate of any server-switch link is a function of the client-side bandwidth. During times of reduced activity or network traffic, a server 111 may be servicing requests from a limited number of clients many of whom may have significant bandwidth limitations. If a server 111 is servicing requests from a single client that is connected to WAN 105 via a conventional modem connection, the effective data rate required of the server is orders of magnitude below the maximum switches maximum capacity. Under such circumstances, the high cost of maintaining a server-switch link at a high operating frequency does not provide any performance benefit since the performance is limited at the client side.

NIC 121 is configured to compare the effective data rate of its network link with the current operating frequency of the link. If the effective data rate (EDR) is materially different than the current link operating frequency, link operating frequency is modified such that the modified frequency is closer to the EDR than the previous operating frequency. If the EDR is determined in block 406 to be substantially lower than the link operating frequency, NIC 121 then determines in block 408 if the link is capable of operating at a lower operating frequency. As discussed previously, NIC 121 and its corresponding switch port are preferably capable of operating at one of multiple operating frequencies. If NIC 121 is not currently operating at its lowest frequency and its effective data rate is substantially below the current operating frequency, NIC 121 is configured to initiate a negotiation with switch 110 that forces (block 410) the link to operate at a lower operating frequency.

In an embodiment where the server-switch links are Ethernet links, NIC 121 may leverage large portions of the standard IEEE 802.3 negotiation protocol to achieve the modification of the operating frequency. Instead of attempting to establish the highest operating frequency accommodated by the link components, the negotiation that occurs in block 410 (referred to as a modification negotiation) responsive to determining that the data rate is well below the operating frequency attempts to achieve an operating frequency that is the lowest possible operating frequency consistent with the most recently determined effective data rate. Thus, NIC 121 may initially indicate a desired operating frequency to switch 110 during the modification negotiation. If switch 110 is capable of operating at the NIC's desired operating frequency, that frequency will become the operating frequency of the link. If switch cannot

accommodate the NIC's desired operating frequency, the negotiation process will resolve the lowest operating frequency accommodated by the link.

The modification negotiations preferably occur at a frequency that is sufficient to adjust to changing loading conditions but not so frequently as to impact performance negatively from excessive negotiation processing. Since, the length of an IEEE 802.3 standard negotiation is on the order of milliseconds, initiating a negotiation even as frequently as every minute should not impose a substantial burden on performance while providing sufficiently frequent modifications to accommodate changes in loading relatively quickly.

In addition to being able to reduce link operating frequency in response to a relatively low level of bandwidth utilization, server cluster 101 is configured to increase link operating frequency in response to relatively high bandwidth utilization. If the server determines in block 406 that the effective data rate is not less than some specified value or some specified ratio of the link's current bandwidth capacity, it may then determine (block 412) whether the effective data rate is above some specified ratio of the link bandwidth capacity. If the effective data rate is more than 90%, for example, of the link's bandwidth capacity, the server may then attempt to increase (block 414) the link operating frequency in a manner analogous to the manner in which the operating frequency is reduced in blocks 408 and 410 as described above. Thus, the server could determine (block 413) whether a higher operating frequency is available and, if so, initiate a modification negotiation to force (block 414) an increased operating frequency. After an increase (or decrease) in operating frequency, server 111 resumes operating and continues to monitor link bandwidth utilization for subsequent changes. In this manner, server 111 is constantly adjusting the link operating frequency to the minimum value required to achieve a desired level of performance thereby reducing energy consumption and heat dissipation characteristic of higher operating frequencies.

It will be apparent to those skilled in the art having the benefit of this disclosure that the present invention contemplates a system and method for conserving energy in a server cluster environment by optimizing the operating frequency of the network links to reflect the current loading. It is understood that the form of the invention shown and described in the detailed description and the drawings are to be taken merely as presently preferred examples. It is intended that the following claims be interpreted broadly to embrace all the variations of the preferred embodiments disclosed